Analysis of Coreference Relations in the Biomedical Literature

Miji Choi[1,2] **Karin Verspoor**[1] **Justin Zobel**[1]
jooc1@student.unimelb.edu.au {karin.verspoor,jzobel}@unimelb.edu.au
[1]Department of Computing and Information Systems, The University of Melbourne
[2]National ICT Australia

Abstract

In this study, we perform an investigation of coreference resolution in the biomedical literature. We compare a state-of-the-art general system with a purpose-built system, demonstrating that the use of domain-specific knowledge results in dramatic improvement. However, performance of the system is still modest, with recall a particular problem (80% precision and 24% recall). Through analysis of features of coreference, organised by type of anaphors, we identify that differentiated strategies for each type could be applied to achieve further improvement.

1 Introduction

The peer-reviewed scientific literature is a vast repository of authoritative knowledge. However, with around 40,000 new journal papers every month, manual discovery or annotation is infeasible, and thus it is critical that document processing techniques be robust and accurate, to enable not only conventional search, but automated discovery and assessment of knowledge such as interacting relationships (events and facts) between biomolecules such as proteins, genes, chemical compounds and drugs. Biological molecular pathways, for example, integrated with knowledge of relevant protein-protein interactions, are used to understand complex biological processes.

Coreference resolution is an essential task in information extraction, because it can automatically provide links between entities, and as well can facilitate better indexing for medical information search with rich semantic information. A key obstacle is the low detection reliability of hidden or complex mentions of entities involving coreference expressions in natural language texts (Kim et al., 2011a; Miwa et al., 2010). Such anaphoric coreference expressions such as pronouns are mostly ignored by event extraction systems, and are not considered as term occurrences in information retrieval systems.

For example, the following passage includes an interacting relation; the *binding* event between the anaphoric mention *the protein* and a cell entity *CD40* is implied in the text. The mention, *the protein*, refers to the specific protein name, *TRAF2*, previously introduced in the same text.

(1) *...The phosphorylation appears to be related to the signalling events ... to be phosphorylated significantly less than the wild-type protein. Furthermore, the phosphorylation status of* ***TRAF2*** *had significant effects on the ability of* ***the protein*** *to bind to CD40, as evidenced by our ...* [PMID:10080948]

In this paper, we investigate the challenges of biomedical coreference resolution, and provide an evaluation of general domain coreference resolution system on biomedical texts. Prior work demonstrated the importance of domain-specific knowledge for coreference (Choi et al., 2014). We extend that work with a detailed analysis of features of coreference relations with respect to the type of the anaphor defined by a previously proposed framework (Nguyen and Kim, 2008), and propose an efficient strategy towards improved anaphoric coreference resolution in the biomedical literature building on that framework.

2 Background

Related Work

For general coreference resolution, several strategies and methodologies have been developed since 1990's. Centering theory was studied based on syntactic information for resolving pronominal expressions (Kehler, 1995), and a framework based

Miji Choi, Karin Verspoor and Justin Zobel. 2014. Analysis of Coreference Relations in the Biomedical Literature. In *Proceedings of Australasian Language Technology Association Workshop*, pages 134−138.

on the Centering theory was developed for the interpretation of pronouns by identifying patterns of coreference (Gordon and Hendrick, 1997).

An unsupervised system was developed to determine coreference links with a collection of rule-based models (Raghunathan et al., 2010), and the system has been extended by (Lee et al., 2011) with additional processes such as mention detection, discourse processing and semantic-similarity processing. The system was developed targeting to the newswire domain, but has been adopted for the clinical domain (Jindal and Roth, 2013; Jonnalagadda et al., 2012; Dai et al., 2012). The rule-based approach has been demonstrated to slightly outperform a machine learning approach for coreference resolution related to treatment, test and person (Jonnalagadda et al., 2012).

Recently, there was a community-wide shared task for coreference resolution in biomedical literature, the Protein Coreference task at BioNLP 2011 (Nguyen et al., 2011). Four out of six participants produced meaningful performance, but the overall performance of those systems was low with the best system (Kim et al., 2011b) achieving F-score=34% (73% precision and 22% recall).

A Framework in the Biomedical Domain

There have been attempts to define characteristics of coreference resolution in the biomedical domain (Gasperin et al., 2007; Gasperin, 2006; Lin et al., 2004; Castano et al., 2002). Pronominal mentions and definite noun phrases (NPs) are regarded as anaphoric references. A framework proposed by Nguyen and Kim (2008) organises anaphoric mentions into categories: Personal pronoun, Demonstrative pronoun, Possessive pronoun, Reflexive pronoun, and Indefinite pronoun. Additionally, antecedents are categorised into an NP or embedded within a larger NP, and by syntactic structure, including NP with a head noun (definite and indefinite), Conjoint NP (with more than one head), Coordinated NP, and NP with restrictive relative clause.

We will demonstrate that by analysing the performance of coreference systems according to these types, we can identify variation in system behaviour that depends on the type of an anaphor of a coreference relation. Our analysis taking advantage of this organisation points to the value of a differentiated treatment of coreference, where applicable rules depend on the specific characteristics of both anaphor and antecedent.

3 Experiment

We compare an existing coreference resolution system, TEES, that uses a domain-specific named entity recognition (NER) module with an existing general system, Stanford CoreNLP, that does not use a domain-specific NER. The aim is to explore how domain-specific information impacts on performance for coreference resolution involving protein and gene entities. The TEES system, which includes a biomedical domain-specific NER component for protein and gene mentions (Björne and Salakoski, 2011), and the Stanford CoreNLP system, which uses syntactic and discourse information but no NER outputs (Lee et al., 2011), are evaluated on a domain-specific annotated corpus.

3.1 Data Sets

We use the training dataset from the task Protein Coreference at BioNLP 2011 for evaluation of existing coreference resolution systems. The annotated corpus includes 2,313 coreference relations, which are pairs of anaphors and antecedents related to protein and gene entities, from 800 Pubmed journal abstracts. Table 1 presents descriptive statistics of the annotated corpus, in terms of the types identified by the coreference framework introduced previously.

Table 1: Statistics of annotations of the gold standard corpus

Anaphor	Relative pronoun	1,256 (54%)
	Pronoun	671 (29%)
	Definite Noun Phrase	346 (15%)
	Indefinite Noun Phrase	11 (0.5%)
	Non-classified	28 (1%)
Antecedent	Including protein	560
	Including conjunction	217
	Cross-sentence	389
	Identical relation	43
	Head-word match	254

3.2 Results

Performance for identification of coreference mentions and relations of each system evaluated on the annotated corpus is compared in Table 2. The Stanford system achieved low performance with F-score 12% and 2% for the detection of coreference mentions and relations respectively, and produced a greater number of detected men-

tions. The TEES system achieved better performance with F-score 69% and 37% for coreference mention and relation levels respectively, but detected a smaller number, reducing system recall.

Our investigation of low performance by each system at the coreference relation level appears in detail in Table 3. Several factors such as lack of domain-specific knowledge (*Including protein* columns), bias towards selection of closest candidate of antecedent (*Pronoun* row for Stanford), limiting analysis to within-sentence relations (*Cross-sentence* column for TEES), syntactic parsing error (*Relative pronoun* row for Stanford), and disregard of definite noun phrase (*Definite NP* row for TEES) have been observed. The main cause, lack of domain-specific knowledge, is explored below.

The annotated corpus contains 560 coreference relations, where anaphoric mentions refer to protein or gene entities previously mentioned in a text. For those coreference relations, the TEES system outperformed the Stanford system by identifying 155 true positives, far more than the 38 identified by the Stanford system, as shown in Table 4.

Table 4: Result of performance of existing systems for coreference relations involving protein names

		Stanford	TEES
Output	(TP)	38	155
	(FP)	1,732	46
Precision (%)		0.02	0.77
Recall (%)		0.07	0.28
F-score (%)		0.03	0.41

The Stanford system also produces a large number of false positives. The Stanford system also produces a large number of false positives. Many of these are coreference relations where an anaphor and an antecedent are identical, or have a common head word (the main noun of the phrase), for example, *IL-2 transcription* (anaphor) – *IL-2 transcription* (antecedent), or *1E8 cells – CD19 cross-linked 1E8 cells*. Such relations are not annotated in the gold standard, and hence are counted as false positives, while they may in fact be linguistically valid coreference relationships. The gold standard defines a different scope for the coreference resolution task than the Stanford system.

On the other hand, the TEES system achieved 77% precision, but still only 28% recall. The main reason for the low recall is that the system is limited to coreference relations where anaphors and antecedents corefer within a single sentence. Even though anaphors mostly link to their antecedents across sentences, the system still identified 155 correct coreference relations by taking advantage of domain-specific information provided through recognition of proteins.

Example 1 above demonstrates how the process of NER in the biomedical domain helps to determine correct coreference relations. The anaphor, ***the protein*** is correctly identified as referring to ***TRAF2*** by the TEES system, but the Stanford system links it to the incorrect antecedent *the wild-type protein* (underlined).

4 Discussion

4.1 Differentiated strategy by anaphor type

We have shown that domain-specific information helps to improve performance for coreference resolution, but the domain-specific system achieved lower recall, with 56% recall of coreference mentions and 24% recall of coreference relations. Features of coreference relations have been analysed following the framework focusing on types of anaphors, but the structure of antecedents have not been considered in this study. Differentiated approaches are considered for each type as shown in Table 5.

Table 5: Differentiated approaches based on anaphor types

Anaphor	**Approaches**
Relative pronoun	Syntactic information
Pronoun	Syntactic information Semantic information
Definite NP	Semantic information Head-word match

Relative Pronouns

As for the type of Relative pronouns, syntactic information results is critical for determining their antecedents. In our analysis, relative pronouns annotated in the gold standard corpus consist of *which*, *that*, and other *wh-* pronouns e.g., *whose*, and *where*. In particular, 100% of *which*, and 75% of *that* are tagged with the WDT Part-of-speech (POS) tag by the Stanford parser. A majority of antecedents for those relative pronouns are mentions placed directly before the relative pronouns,

Table 2: Results of evaluation of existing systems on the annotated corpus

	Stanford		TEES	
	Mentions	Relations	Mentions	Relations
Gold annotation	4,367	2,313	4,367	2,313
System detected	12,848	7,387	2,796	707
Exact match	1,006	112	2,466	564
Precision (%)	0.08	0.02	0.88	0.80
Recall (%)	0.23	0.05	0.56	0.24
F-score (%)	0.12	0.02	0.69	0.37

Table 3: Analysis of performance of existing systems comparing to the annotated corpus

		Stanford				TEES			
		Cross-sentence	Internal-sentence	Including protein	Including conjunction	Cross-sentence	Internal-sentence	Including protein	Including conjunction
Relative pronoun	TP	0	1	0	0	0	393	116	9
	FP	0	2	1	0	0	86	27	4
Pronoun	TP	7	62	28	10	0	162	37	9
	FP	675	302	197	132	0	47	15	2
Definite NP	TP	35	7	10	1	0	7	2	1
	FP	1,183	194	483	179	0	3	1	0
Nonclassified	TP	0	0	0	0	0	1	0	0
	FP	4,129	650	1,187	632	0	5	3	0

or close to the relative pronouns within a 10 character span in the same sentence. However, this approach has a defect that it would fail to find correct antecedents, if texts are incorrectly parsed by the syntactic parser. There are 63 *that* tokens tagged with the DT, that should be labelled with WDT.

Definite Noun Phrases

In the gold standard corpus, there are 127 out of 346 coreference relations where an anaphor that is a Definite NP has a biomedical named-entity as an antecedent, and 176 of the 346 anaphors share head-words e.g., *genes*, and *proteins* with the antecedent. For the relations neither involving proteins nor with shared head-words, other approaches are applied, such as Number Agreement to coreference relations e.g., *these genes* is plural and so must refer to multiple genes – *actin and fibronectin receptor mRNA* – and (domain-specific) Semantic Knowledge, such as the similarity between two terms. For example, "complex" and "region" in the coreference relation *the binding complexes – this region of the c-myb 5' flanking sequence* must be recognised as (near-) synonyms.

Pronouns

Pronouns are a more difficult type of anaphor to resolve than others, because they do not include helpful information to link their antecedents. In our data, the resolution scope of antecedents for Subject pronouns is defined within the previous sentence, while the Non-subject pronouns can refer anywhere in the text. For the Non-subject pronouns, semantic information based on its context is important. Among 238 coreference relations where an anaphor is a Pronoun, and their antecedents embed one or more specific protein names, 191 include protein-relevant words (defined by (Nguyen et al., 2012)), such as *binding, expression, interaction, regulation, phosphatase, gene, transactivation, transcription*.

5 Conclusion

In this study, we have explored how domain-specific knowledge can be helpful for resolving coreferring expressions in the biomedical domain. In addition, features of coreference relations have been analysed focusing on the framework of anaphors. By taking advantages of the framework, we expect that differentiated approaches for each type of anaphors will improve the task of coreference resolution, and further investigation according to antecedent types with syntactic characteristics is being left for future work.

Acknowledgments

This work was supported by the University of Melbourne, and by the Australian Federal and Victorian State governments and the Australian Research Council through the ICT Centre of Excellence program, National ICT Australia (NICTA).

References

Jari Björne and Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 183–191. Association for Computational Linguistics.

José Castano, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature.

Miji Choi, Karin Verspoor, and Justin Zobel. 2014. Evaluation of coreference resolution for biomedical text. *MedIR 2014*, page 2.

Hong-Jie Dai, Chun-Yu Chen, Chi-Yang Wu, Po-Ting Lai, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2012. Coreference resolution of medical concepts in discharge summaries by exploiting contextual information. *Journal of the American Medical Informatics Association*, 19(5):888–896.

Caroline Gasperin, Nikiforos Karamanis, and Ruth Seal. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of DAARC*, volume 2007. Citeseer.

Caroline Gasperin. 2006. Semi-supervised anaphora resolution in biomedical texts. In *Proceedings of the HLT-NAACL BioNLP workshop on linking natural language and biology*, pages 96–103. Association for Computational Linguistics.

Peter C Gordon and Randall Hendrick. 1997. Intuitive knowledge of linguistic co-reference. *Cognition*, 62(3):325–370.

Prateek Jindal and Dan Roth. 2013. Using domain knowledge and domain-inspired discourse model for coreference resolution for clinical narratives. *Journal of the American Medical Informatics Association*, 20(2):356–362.

Siddhartha Reddy Jonnalagadda, Dingcheng Li, Sunghwan Sohn, Stephen Tze-Inn Wu, Kavishwar Wagholikar, Manabu Torii, and Hongfang Liu. 2012. Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules. *Journal of the American Medical Informatics Association*, pages amiajnl–2011.

Andrew Kehler. 1995. *Interpreting cohesive forms in the context of discourse inference*. Ph.D. thesis, Citeseer.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011a. Overview of bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 1–6. Association for Computational Linguistics.

Youngjun Kim, Ellen Riloff, and Nathan Gilbert. 2011b. The taming of reconcile as a biomedical coreference resolver. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 89–93. Association for Computational Linguistics.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The genia event extraction shared task, 2013 edition-overview. *ACL 2013*, page 8.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics.

Yu-Hsiang Lin, Tyne Liang, and T Hsinehu. 2004. Pronominal and sortal anaphora resolution for biomedical literature. In *ROCLING*.

Makoto Miwa, Rune Sætre, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. Event extraction with complex event classification using rich features. *Journal of bioinformatics and computational biology*, 8(01):131–146.

Ngan LT Nguyen and Jin-Dong Kim. 2008. Exploring domain differences for the design of pronoun resolution systems for biomedical text. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 625–632. Association for Computational Linguistics.

Ngan Nguyen, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Overview of the protein coreference task in bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 74–82. Association for Computational Linguistics.

Ngan Nguyen, Jin-Dong Kim, Makoto Miwa, Takuya Matsuzaki, and Junichi Tsujii. 2012. Improving protein coreference resolution by simple semantic classification. *BMC bioinformatics*, 13(1):304.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.